12-10-2017

Reinforcement Learning for Profit Maximization of Recommender Systems

Jo Yong Ju
KyungHee University, young456@khu.ac.kr

Il Young Choi
KyungHee University, choice102@khu.ac.kr

Hyun Sil Moon
KyungHee University, pahunter@khu.ac.kr

Jae Kyeong Kim
KyungHee University, jaek@khu.ac.kr

Reinforcement Learning for Profit Maximization of Recommender Systems

Completed Research Paper

Jo Yong Ju
KyungHee University
young456@khu.ac.kr

Il Young Choi
KyungHee University
choice102@khu.ac.kr

Hyun Sil Moon
KyungHee University
pahunter@khu.ac.kr

Jae Kyeong Kim
KyungHee University
jaek@khu.ac.kr

Introduction

Since the 1990s, various recommender systems have been developed and become more important by increasing data size in many fields such as music, online video, and book (Bobadilla et al., 2013: Choi et al., 2012; Choi et al., 2016; Lee et al., 2010). Recently, recommender systems have also been studied in applications such as Internet of Things (IoT), social network, and so on.

Many researches on the recommender systems have focused on the accuracy of recommendations because the recommender systems help customers to find the products which are suited to their preferences. However, recommendation service which do not meet customer expectations may lead to rejection of the recommendation and even contempt for the recommendation service (Fitzsimons & Lehmann, 2004). Accordingly, recent studies have been also conducted in terms of other performance metrics such as diversity and stability.

However, one of the important aspect of business is to increase the profit margins of the company through recommender systems. That is, the main purpose of the recommender system is to increase the profits of the company by providing information that customers would like. Thus, some studies have tried to find the model which earns more profits (Azaria et al., 2013; Chen et al., 2008; Das et al., 2010; Lu et al., 2014). However, it is not easy to directly estimate the increase of profits. Moreover, even though there is a significant profit gap whether the company applied to the recommender systems or not, there are few studies which demonstrate higher profits.

In this study, we propose a reinforcement learning-based recommendation methodology for increasing the profits of the company which considers the price of the products. In order to apply reinforcement learning to recommender systems, we regard the price of products as the reward of reinforcement learning. In other words, we assume that the company will recommend a product that is highly profitable in the long term, and it is designed to learn about not only pricing factors, but also the future earnings of the products. Although there are many techniques for reinforcement learning, we use the Markov Decision Process(MDP). The Markov property means that the probability of future state are determined solely by the present state. Therefore, the proposed methodology conducts recommendations based on both purchased products at a specific point of sale and the transfer probability of purchasing at the next time. It is expected that a greater margin of profit will be obtained from the extent that there is no significant difference in accuracy, compared with the traditional recommendation model. Therefore, we expect that our methodology will earn more profit than the traditional recommender systems.

Related Work

Reinforcement Learning

Reinforcement learning has been studied early in the field of cybernetics, statistics, psychology, and neurological science (Kaelbling et al., 1996). The Reinforcement learning methods are distinguished from typical models such as supervised learning and unsupervised learning because they are defined by learning how to study problems, rather than learning how to study methods (Sutton & Barto, 1998). The methods of reinforcement learning are based on several studies. The motive of reinforcement learning has become a case of psychologist Skinner's Box experiments. The Skinner's reinforcement theory is that it pays the highest rewarding behavior based on past compensation, which led to becoming a keynote for reinforcement learning. Another basis of reinforcement learning is from the bellman's optimal control studies in the theoretical aspects. Bellman solved the optimal control problem through the bellman equation, and furthermore created the Markov decision process to complete the foundation of reinforcement learning. Skinner's Reinforcement learning and bellman's Markov decision process merged into Sutton's temporal difference (Sutton, 1988) and Watkins Q-learning (Watkins & Dayan, 1992). Recently, the development of reinforcement learning leads to studies such as temporal-difference search in computer Go (Silver et al., 2012), game control AI (Mnih et al., 2015), and Deep mind-AlphaGo (Silver et al., 2016). The latest research's results showed that Google's AlphaGo which is based on the deepQ learning has won the game with the man.

Most reinforcement learning researches are based on the formalism of Markov decision process (Puterman, 2014). Although reinforcement learning is by no means restricted to Markov decision process, reinforcement learning's basis is motivated from Markov decision process such as discrete time, countable state and action formalism (Barto & Mahadevan, 2003). The reinforcement learning agent gets rewarded by acknowledging the environment and responding to the environment. Accordingly, reinforcement learning discovers an optimal action through trial-and-error interactions with its environment (Kober et al., 2013). Recently, several studies have proposed recommendation methodologies based on reinforcement learning techniques (Gaeta et al., 2016; Mahmood & Ricci, 2007; Mahmood & Ricci, 2009; Taghipour et al., 2007; Wang et al., 2014).

Profit maximization on Recommender Systems

The recommendation system can enhance the corporate expectation by showing the customer preference products. So, many studies on recommendation systems have been focused and developed to encourage accuracy and to provide various recommendations. But, the previous studies cannot increase profits directly, only indirectly.

Without considering the customer's preference, recommendations that encourage the product's profit may result in making customers purchase with curiosity. But, if the customer continues to be exposed, the customer will eventually disregard the recommendations. Accordingly, some studies have proposed the models considering both customer's preference and the revenue of the firm (Chen et al., 2008; Das et al., 2010). Das et al.(2010) have used the concept of trust to the customer to define the relationship between the customer's preference and the duration of the trust relationship. Any recommendations based on trust can yield higher expectations. Chen et al.(2008) have proposed methodology which integrated price profitability into the traditional recommender systems.

The companies can increase profits by simply recommending a highly priced commodity. However, if the trust of customer toward the companies continues to decline, the efficiency and performance of the recommendations will gradually decrease (Das et al., 2010). In other words, recommendation methodology for profitable one-shot sales may lose customers' trust.

In this study, we propose a methodology that enables the companies to get higher returns on sales for long periods of time. Especially, the proposed methodology does not exist a trade-off between the accuracy of recommendations and the companies' profit margins because of considering the price of products and the association with the products purchased from the previous records.

Methodology

Overview

Many recommendation methods have been studied in order to better predict the accuracy of the customer's preferences. These methods have proven that there is a positive correlation between sales volume and the accuracy. However, there are few studies to propose the recommendation methods for increasing profit though sales volume is important in aspect of business.

In this study, we develop a methodology to recommend products which can earn maximum profits in long-term periods. The methodology uses the user's purchase data and pricing factors. In this paper, we use the future price by modifying Q-learning as reward of state-action value. Overall procedure of the proposed methodology is shown in Figure 1.

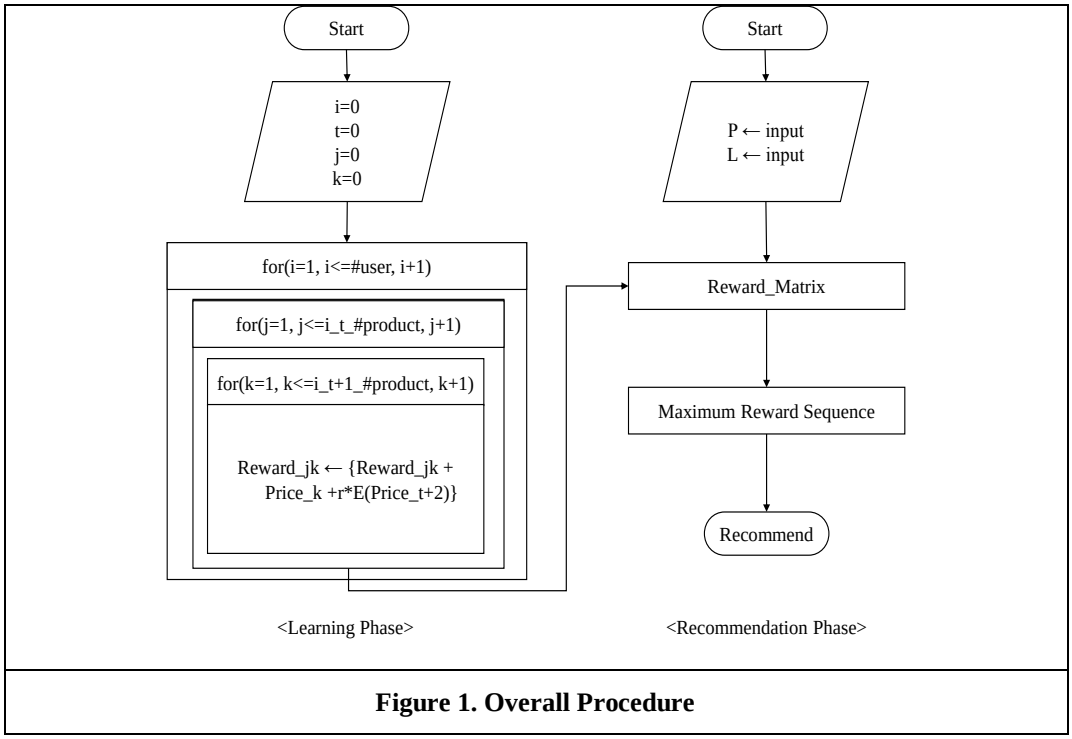

Figure 1. Overall Procedure

In the learning phase, the model preprocesses data for learning process and it learns the state-action reward matrix using reinforcement learning. The model provides recommendations based on the learned state-action reward matrix and it decides the policy which the reward is maximum at each state. In the recommendation phase, the model recommends the product whose reward is maximum based on the product which was bought most recently.

In this study, let $I = \{i_1, i_2,..., i_{n-1}, i_n\}$and $U = \{u_1, u_2,..., u_{m-1}, u_m\}$ be a set of products and a set of users, respectively. The tuples of Markov decision process are defined as the following criteria for applying the product recommendation of the transaction data.

· State: Set of purchased products at time t with user u_i

$S = \{s_1, s_2,..., s_{n-1}, s_n \mid 1 \le k \le n, s_k \in I\}$

· Action: Set of purchased product at time t+1 with user u_i

$A = \{a_1, a_2,..., a_{n-1}, a_n \mid 1 \le j \le n, a_j \in I\} \subset I$

· State transition probability: $P(s_{t+1} = i_{n+1} | s_1 = i_1, s_2 = i_2, ..., s_t = i) = P(s_{t+1}=i_{n+1} | s = i)$

· Discount factor: Discount factor is a degree of reduced present value of the future rewards. We define discount factor as interval value in $\gamma \in [0, 1]$ through experiment.

· Reward: Reward is defined on the profit basis because the price is the profit of selling products, the equation $R(s, a)$ = price of product 'a' can be formed.

The proposed system learns the model using Markov decision process' five tuples and recommends user with the product whose $Q(s, a)$ is maximum.

Learning Phase

As explained above, it is required to preprocess the transaction data for the proposed methodology. The order of state and action must be defined because reinforcement learning of the proposed methodology is based on Markov decision process.

So, dividing and arranging the transaction data according to the purchase orders are required. We rearrange the purchase order for each customer and convert the date back to the order of purchase such as Table 1. Here, *u*, *t*, and *i* mean a user, time, and a product, respectively.

Table 1. An Example of Purchasing Matrix

	t_1	t_2	t_3	...
u_1	i_1	i_4	i_3	...
u_2	i_3	i_4	...	...
u_3	i_2	i_3	...	...
u_4	i_1	i_3	...	...

The table 1 represents the learning information of the model. For example, if i_1 at time t_1 is the state of u_1, i_4 at time t_2 will become next action of u_1. The reward $R(s,a)$ given at this moment can be 2.5($) if price of i_4 is 2.5$. So, the equation of learning Q-learning can be shown following as:

$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + R(s_t, a_t) + \gamma(\sum_{a_{t+1}} P_{s_{t+1} a_{t+1}} * R(s_{t+1}, a_{t+1}))$, where Q: s x a →R, and P is probability transition.

At state s_t and agent action a_t, agent can get reward $R(s_t, a_t)$ that is price of the product a_t, and then $Q(s_t, a_t)$ is updated by reward $R(s_t, a_t)$. Moreover, $Q(s_t, a_t)$ is added by future discounted reward. We define the terminated sequence as two step of states. If sequence is terminated sequence as one step, the model cannot learn the future reward. But if the sequences are terminated as more two steps, the model's learning time is much longer than that of *(the number of procusts)*^{2}.

So, Q(s, a) matrix, which represents the maximal reward at the state, is learned through Q-learning. The matrix has accumulated reward of each state and action. Therefore, *Q(s, a)* has two characteristics. First, the higher the frequency of the product, the higher the value of the *Q(s, a)*. If a particular item a_j is bought frequently after purchasing a product of a specific state s_k, that is an important factor in increasing the value of $Q(s_k, a_j)$. Second, The higher the price of the product, the higher the *Q(s, a)*. If the price of a product is high, it is possible to obtain a high *Q(s, a)*. This means that the current state of the agent is more likely to transfer to a higher price. When recommending, this action becomes the action which has the highest *Q(s, a)* values in accordance with the state.

Recommendation Phase

The learning of the model terminates when the customer's purchasing is over. And then, recommendation

is offered. That is, the model identifies the state s_k of the customer and locates a_j that is the highest value of $Q(s_k, a_j)$ in the $Q(s, a)$ matrix. Therefore, a_j will be recommended to the target customer as following.

$$Rec(s_k) = argmaxQ(s_k, a_j) \ \ (where, 1 \leq j \leq max\ j)$$

In this phase, the purchased product is removed from the list of recommendation. And the number of the recommended product is same to the number of the purchased product in the most recent time. For example, if u_1 bought three products at t_1, the number of recommended products at t_2 will be three.

Experimental Results

Data and Experimental setup

For our experiments, dataset is obtained from Dunnhumby (https://www.dunnhumby.com/sourcefiles). This dataset contains transactions of 2,500 households who are frequent shoppers at a retailer. We analyzed only 100 customers with a high frequency because lower frequency reduces the rate of learning to be learned by data. The selected dataset contains 2,069 products and 417, 949 transactions.

Due to the characteristic of retail data, the accuracy of the model is affected by the factors and time. In order to remove the elements of time, this study uses the leave-p-out validation method. The leave-p-out validation method can control the accuracy of the data because the accuracy of models can be changed by seasonal factors (Celisse and Robin, 2008; Moon et al., 2013). In this study, we first divide the entire data into 10 time periods. And then we select the model to measure the performance by using one of the following data. For example, if p is 6, we can use the data from the 1st to 6th of the entire data slices to study the model, and then measure the performance with the 7th data. And we repeat this learning until the evaluation with the 10th data.

To compare the performance between our methodology and Markov decision process-based model as a benchmark system, we use two metrics such as the expected value and the real sales value. These metrics are described as follows:

$$\text{Expected Value} = \frac{\sum price\ of\ recommended\ products}{number\ of\ recommended\ products}$$

$$\text{Real Sales Value} = \frac{\sum price\ of\ purchased\ products\ with\ recommended}{number\ of\ purchased\ products\ with\ recommended}$$

Results

Before the comparison of the performance for each recommender system, we set the discount factor as 0.5 because it is best served in terms of accuracy and expected value. Figure 2 shows the differences of the expected value between the proposed methodology and the Markov decision process -based model.

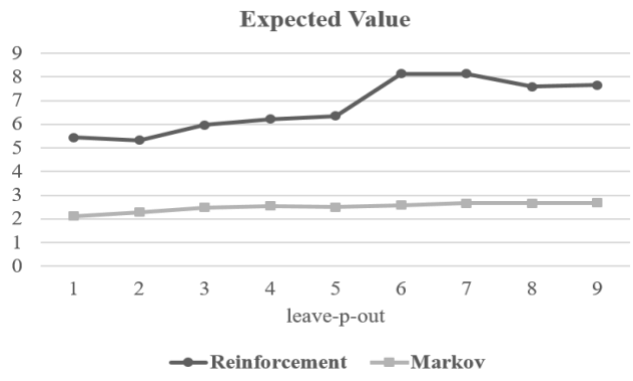

Figure 2. Comparison of Expected Value

As shown in Figure 2, the expected value of the proposed methodology is approximately 3 times larger than the benchmark. That is, our methodology can recommend products with higher prices than the benchmark. However, it does not mean that the company's profits are also higher. Therefore, we next compare the real sales value in Figure 3.

Figure 3. Comparison of Real Sale Value

As a result, there are not significant differences between Figure 2 and 3. Especially, in Figure 3, our methodology can increase the company's profit about 1.5 times. Therefore, we can conclude that our methodology can provide a way to induce more profit margins from the corporate sector and increase corporate profits. For a more detailed test, we performed the non-parametric Mann–Whitney statistical test. The results indicated that the performance of the proposed system was significantly higher than that of the benchmark system, as shown in Table 2.

Table 2. Results of Mann–Whitney Test

(a) Expected Value

	N	Mean Rank	Mann-Whitney U	Wilcoxon W	Z	Exact Sig.

Reinforcement	9	14	0.000	45	-3.578	0.00**
Markov	9	5				
** p<0.05						
(b) Real Sales Value						
	N	Mean Rank	Mann-Whitney U	Wilcoxon W	Z	Exact Sig.
Reinforcement	9	14	0.000	45	-3.576	0.00**
Markov	9	5				
** p<0.05						

Conclusion and Future Works

In traditional studies, the recommender systems help customers to find the products which is suitable to their preferences. The customers broaden the purchase pattern by recommendation because the recommendation list is different to the previous purchased products. Accordingly, the recommender systems can be applied as a marketing strategy that allows businesses to make higher profits as companies provide recommendations. Therefore, the revenues from businesses should be considered when providing recommendations. In order to solve this problem, we focus on the long-term earnings which provide a maximum incentive from the standpoint of offering recommendations.

Through experiments, we find that the proposed methodology shows improved outcomes in terms of anticipated earnings. That is, the perceived value of the proposed methodology is better than that of the existing recommender systems. Moreover, the proposed methodology can be used for real-time recommender system. Although data accumulates in real-time, traditional recommender systems should repeat train process from the beginning to use new data. However, our methodology can be updated immediately because it only learns from newly created data.

However, our study has some limitations. First, we estimated the performance related with profits. Therefore, we will estimate the accuracy of our methodology and try to find a way to increase the performance related with both accuracy and profits. Second, recently, some studies try to use a variety of data such as unstructured data. Therefore, we will consider additional data for our methodology to improve the performance.

References

Azaria, A., Hassidim, A., Kraus, S., Eshkol, A., Weintraub, O., and Netanely, I. 2013. "Movie Recommender System for Profit Maximization," in *Proceedings of the 7th ACM conference on Recommender systems*, pp. 121-128.

Barto, A. G., and Mahadevan, S. 2003. "Recent Advances in Hierarchical Reinforcement Learning," *Discrete Event Dynamic Systems* (13:4), pp. 341-379.

Bobadilla, J., Ortega, F., Hernando, A., and Gutiérrez, A. 2013. "Recommender Systems Survey," *Knowledge-based systems* (46), pp. 109-132.

Celisse, A., and Robin, S. 2008. "Nonparametric Density Estimation by Exact Leave-p-out Cross-validation," *Computational Statistics & Data Analysis* (52:5), pp. 2350-2368.

Chen, L. S., Hsu, F. H., Chen, M. C., and Hsu, Y. C. 2008. "Developing Recommender Systems with the Consideration of product profitability for sellers," *Information Sciences* (178:4), pp. 1032-1048.

Choi, I. Y., Oh, M. G., Kim, J. K., & Ryu, Y. U. 2016. "Collaborative Filtering with Facial Expressions for Online Video Recommendation," *International Journal of Information Management* (36:3), pp. 397-402.

Choi, K., Yoo, D., Kim, G., and Suh, Y. 2012. "A Hybrid Online-product Recommendation System: Combining Implicit Rating-based Collaborative Filtering and Sequential Pattern analysis," *Electronic Commerce Research and Applications* (11:4), pp.309-317.

Das, A., Mathieu, C., and Ricketts, D. 2010. "Maximizing Profit using Recommender Systems," in *Proceedings of the International Conference on World Wide Web*.

Fitzsimons, G.J. and Lehmann, D.R. 2004, "Reactance to Recommendations: When Unsolicited Advice Yields Contrary Responses", *Marketing Science* (23:1), pp. 82-94.

Gaeta, M., Orciuoli, F., Rarità, L., and Tomasiello, S. 2016. "Fitted Q-iteration and Functional Networks for Ubiquitous Recommender Systems," *Soft Computing*, pp. 1-9.

Kaelbling, L. P., Littman, M. L., and Moore, A. W. 1996. "Reinforcement Learning: A Survey," *Journal of Artificial Intelligence Research* (4), pp. 237-285.

Kober, J., Bagnell, J. A., and Peters, J. 2013. "Reinforcement Learning in Robotics: A Survey," *The International Journal of Robotics Research* (32:11), pp. 1238-1274.

Lee, S. K., Cho, Y. H., and Kim, S. H. 2010. "Collaborative Filtering with Ordinal Scale-based Implicit Ratings for Mobile Music Recommendations," *Information Sciences* (180:11), pp. 2142-2155.

Lu, W., Chen, S., Li, K., and Lakshmanan, L. V. 2014. "Show Me the Money: Dynamic Recommendations for Revenue Maximization," in *Proceedings of the VLDB Endowment*, pp. 1785-1796.

Mahmood, T., and Ricci, F. 2007. "Learning and Adaptivity in Interactive Recommender Systems," in *Proceedings of the ninth international conference on Electronic commerce*, pp. 75-84.

Mahmood, T., and Ricci, F. 2009. "Improving Recommender Systems with Adaptive Conversational Strategies," in *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pp. 73-82.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... and Petersen, S. 2015. "Human-level Control through Deep Reinforcement Learning," *Nature* (518:7540), pp. 529-533.

Moon, H. S., Kim, J. K., and Ryu, Y. U. 2013. "A Sequence-based Filtering Method for Exhibition Booth Visit Recommendations," *International Journal of Information Management* (33:4), pp. 620-626.

Puterman, M. L. 2014. Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... and Dieleman, S. 2016. "Mastering the Game of Go with Deep Neural Networks and tree search," *Nature* (529:7587), pp. 484-489.

Silver, D., Sutton, R. S., and Müller, M. 2012. "Temporal-difference Search in Computer Go," *Machine Learning* (87:2), pp. 183-219.

Sutton, R. S. 1988. "Learning to Predict by the Methods of Temporal Differences," *Machine Learning* (3:1), pp. 9-44.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction* (1:1), Cambridge: MIT press.

Taghipour, N., Kardan, A., and Ghidary, S. S. (2007, October). Usage-based Web Recommendations: a Reinforcement Learning Approach. In Proceedings of the 2007 ACM conference on Recommender systems, pp. 113-120.

Wang, X., Wang, Y., Hsu, D., and Wang, Y. 2014. "Exploration in Interactive Personalized Music Recommendation: a Reinforcement Learning Approach," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (11:1), 7.

Watkins, C. J., and Dayan, P. 1992. "Q-learning," *Machine Learning* (8:3-4), pp. 279-292.